DOCUMENT RESUME

ED 318 760 TM 014 802

AUTHOR Ferrara, Steven; Walker-Bartnick, Leslie
TITLE Detecting and Analyzing Differential Item Functioning in an Essay Test Using the Partial Credit Model.
PUB DATE Apr 90
NOTE 33p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 17-19, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; Data Analysis; *Essay Tests; Evaluation Methods; Grade 9; Graduation Requirements; High Schools; *High School Students; *Item Bias; *Item Response Theory; Research Methodology; *Scoring; Standardized Tests; Testing Problems; *Writing Evaluation
IDENTIFIERS Maryland Writing Test; *Partial Credit Model

ABSTRACT

A procedure was developed to detect differential item functioning (DIF) in a standardized essay test using the Partial Credit Model, the general polychotomous form of the Rasch model. Using a panel of experts in the writing process, hypothesized explanations for DIF at some score points were developed. Data for the study included averaged narrative and explanatory essay scores on the Maryland Writing Test for over 47,000 ninth-grade students from the 1989 annual administration. Student score records were randomly selected to create separate data files of about 1,200 White, Asian, Black, Hispanic, male, and female students each. DIF was investigated using a statistical procedure and via an expert panel review procedure. The statistical procedure involved calibrating prompts in separate subgroups and calculating Draba's DIF statistic. The panel review process and results appear to be reasonably helpful for prompt development, but less useful at the post-administration prompt analysis. Implications for research are discussed, and recommendations are made for applications of both procedures in other contexts. Five tables and two graphs summarize study findings. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

STEVEN FERRARA

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED318760

Detecting and Analyzing Differential Item Functioning in an Essay Test Using the Partial Credit Model

Steven Ferrara
Leslie Walker-Bartnick

Maryland State Department of Education

Paper presented at the annual meeting of the National Council on Measurement in Education, April 19, 1990, Boston

Our thanks to Joseph Bellino, Susan Jablinske, Leslie Jones, and Portia White for their contributions to this study

TM014902

Detecting and Analyzing Differential Item Functioning in an Essay Test Using the Partial Credit Model

Steven Ferrara
Leslie Walker-Bartnick

Maryland State Department of Education
April 11, 1990

In recent years a great deal of measurement research has focused on various procedures for detecting differential item functioning (DIF) in test items (see, for example, Cole & Moss, 1989, pp. 208-212). This research has included comparing DIF detection procedures in simulated data with known degrees and patterns of DIF and in real data from existing tests (e.g., Baghi & Ferrara, 1989, 1990). Research on the feasibility, reliability, and validity of various DIF detection procedures has been limited to data from dichotomously scored items (see Benson, 1987). Further, this research has focused primarily on detecting DIF, not the consequences of various responses to detected DIF.

In addition, in recent years much research and discussion has focused on the development and technical characteristics of standardized essay tests (e.g., Breland, Camp, Jones, Morris, & Rock, 1987; Ferrara, 1987; Goldberg & Walker-Bartnick, 1988). Some of this discussion and research includes evaluation of use of the Partial Credit

model of Masters (1982; described below) to equate intact forms of essay tests (Masters & Beswick, 1986; Phillips, 1987); equate direct with indirect writing assessments (Phillips, Mead, & Ryan, 1983); construct narrative writing tasks (Harris, Laan, & Mossenson, 1988); calibrate several types of writing tasks (Pollitt & Hutchinson, 1987); and to construct an equated essay prompt bank (Ferrara & Walker-Bartnick, 1989).

Earlier, we proposed that the Partial Credit model could be used to investigate DIF in essay tests (Ferrara & Walker-Bartnick, 1989). Our position in this study is that (a) theoretical and practical concerns for fairness, reliability, and validity of tests comprised of dichotomously scored items apply equally to polychotomously scored assessments, including essay tests; and (b) methods and procedures for detecting DIF in such assessments should be developed and evaluated.

In this study we propose to (a) develop and demonstrate a procedure for detecting DIF in a standardized essay test using the Partial Credit model; and (b) develop, in a panel of educators who are experts on the writing process, hypothesized explanations for DIF that can occur at some, but not all, score points in polychotomously scored essays.

Background

Differential Item Functioning

The phenomenon of test items that function differently for different examinee subgroups -- that is, that may be easier or more difficult for one group than they are for another -- is usually referred to as differential item functioning (DIF). Test items may be considered to be functioning differentially -- that is, "biased" -- when a focus group scores disproportionally higher or lower than a reference group as a result of factors irrelevant to the ability or skills domain being measured. When using item response theory (IRT) to detect DIF, an item is considered "unbiased" if the item characteristic curves (ICCs) from calibration of an item in two compared subgroups are the same (Crocker & Algina, 1986, p. 377; Ironson, 1982, p. 118). In this study we use the IRT definition for DIF and compare Rasch model difficulty values to determine the existence of potential DIF.

Throughout this paper we refer to items that may (a) function differentially for compared subgroups, (b) be "biased" against a subgroup, (c) "favor" one subgroup over another, and that may be (d) easier or more difficult for a particular subgroup. In all such references we intend to portray the notion that an item may function differently in two examinee groups of equal ability.

The Partial Credit Model

The Partial Credit (PC) model is the general polychotomous form of the Rasch model. It is designed for use with personality, attitude, educational achievement, and other items that elicit responses that are scored in ordered categories; that is, for item responses that can be awarded partial credit and not scored only correct/incorrect. Unlike the Rasch model for dichotomous items, items calibrated with the PC model have associated with them several difficulty values, each indicating the difficulty of reaching a "step" along the way to successful completion of an item. Like the Rasch model for dichotomous items, ability estimates, item and test characteristic curves and information functions, fit statistics, and standard errors at each ability level can be calculated and used for item and test evaluation.

Method and Procedures

Instrument

The Maryland Writing Test (MWT) is administered annually to students in grades 9-12. Students must pass this test and multiple-choice tests in reading, mathematics, and citizenship in order to receive a Maryland high school diploma. The MWT is comprised of two essays written in response to a narrative and an explanatory prompt. Each

prompt consists of three paragraphs which identify the elements of the writing task. Paragraph one identifies the topic, purpose, audience, and form (e.g., friendly letter) for the writing task; paragraph two contains planning suggestions directly relevant to the elements in paragraph one; and paragraph three restates the task in a single sentence. Table 1 contains the first paragraph from the narrative and explanatory prompts from the 1989 MWT.

Table 1

Prompts from the 1989 Maryland Writing Test (First Paragraph Only)

Narrative
Suppose a friend asks you to write about a time you were angry. This may have happened recently or long ago. Write a letter to a friend telling about a time you were angry.
Explanatory
Suppose a teacher has asked you to explain an activity you know how to do. This might be an activity in which you make something, fix something, play something, or do something else. Write a paragraph or more for your teacher explaining an activity you know how to do.

Each essay is scored independently by two trained raters who receive packets of 25 randomly sorted essays. Raters are trained in application of Maryland's scoring rubrics and modified holistic scoring procedures. Raters assign scores of 1, 2, 3, or 4 to essays according to the rubric. Pairs of scores for each essay are averaged

(discrepant scores are resolved by a third independent rating), and pairs of averaged essay scores are summed to produce each student's total test score. The total test scale contains 16 score points ranging from 0.0 to 8.0 in half-point units. The narrative and explanatory score scales each contain seven score points ranging from 1.0 to 4.0 in half-point units. The feasibility of using the PC model with MWT essay data has been demonstrated in several previous studies (e.g., Ferrara & Walker-Bartnick, 1989).

Subjects and Data Set

The data for this study are averaged narrative and explanatory essay scores of over 47,000 grade 9 students from the 1989 annual administration. These students probably participated in intensive writing instruction in preparation for the test and in sustained writing instruction for several years prior to grade 9. Student score records were randomly selected to create separate data files of approximately 1,200 White, Asian, Black, Hispanic, male, and female students each.

Procedures

Two sets of procedures were used to investigate DIF in this essay test: (a) a statistical procedure, to detect differentially functioning steps within essay prompts; and (b) an expert panel review procedure, to develop

hypothesized explanations of any score scale steps that may be identified as functioning differentially.

Statistical detection procedure. The purpose of this analysis was to identify score categories (actually, steps between each category) that are functioning differentially for one subgroup over another; that is, to detect or "flag" DIF in an essay prompt. Statistical detection of DIF in essay prompts was accomplished in two steps. First, essay scores in each of the six data files were calibrated separately using the MSTEPS program (Wright, Congdon, & Schultz, 1988). Second, differentially functioning prompt steps were identified using Draba's statistic (Huynh & Gleaton, 1987; see also Crocker & Algina, 1986, p. 380). Draba's statistic is calculated using compared step difficulty values and their calibration standard errors; that is,

$$B = \frac{d_1 - d_2}{(SE_{d1}^2 + SE_{d2}^2)^{1/2}},$$

where d_1 is a step difficulty value for a reference group, d_2 is the difficulty value for the same step in a focus group, and SE_{d1}^2 and SE_{d2}^2 are corresponding squared standard errors. The statistic is assumed to follow approximately a normal distribution with zero mean and unit standard deviation. Statistical significance is determined

for each calculated value by comparison to a critical value in a standard normal curve table.

Draba's DIF statistic was calculated for each subgroup comparison at each step difficulty using the White subgroup as the reference group for the Asian, Black, and Hispanic subgroups and males as the reference group for females.

Expert panel review procedure. All essay prompt steps that were flagged in the statistical procedure were analyzed by a panel of educators who are expert in writing instruction and assessment and consciously sensitive to race, ethnicity, gender, language, and culture issues relevant to assessment of writing. The panel included four writing experts from Maryland public schools who have considerable experience in direct assessment of writing and in various phases of development of the MWT, and the two authors of this paper. The writing experts include a writing assessment specialist from the Maryland State Department of Education, a secondary English language arts teacher in a predominantly minority school, a member of the Maryland Functional Testing Program item Bias Review Committee, and a secondary English language arts teacher who teaches non-English speaking students; they are, respectively, two black females, an Asian female, and a white male. The purpose of the panel review was to develop hypothesized explanations for differential functioning at

particular steps in an essay prompt for or against a subgroup.

The panel members met for two half-day sessions. Members were introduced to the notion of test item "bias," the more general notion of test bias, and the fundamentals of item response theory; shown the text of differentially functioning multiple-choice items with either "explainable" or "non-explainable" DIF; and shown the results of the statistical analysis in this study. Panel members were then asked to work individually, in pairs, and then with the entire panel to develop hypothesized explanations for the differentially functioning steps contained in Table 2 below, one result at a time. They were asked to consider prompt topic and language; differences in writing instruction programs, school size and location, and community socioeconomic status; differences in rhetorical strategies used in essays (e.g., relating a story or process within an explanation) and language usage (e.g., idiom, vocabulary, grammar); that the result may be "non-explainable"; and other hypothesized explanations for flagged steps. They were also encouraged to pose potential explanations that they felt were reasonably plausible and that were relevant to the specific student subgroup comparison and score scale step at which DIF had been detected.

Panel members were asked to develop hypothesized explanations for two sets of intuitively expected results first (i.e, for potential DIF that "favors" females over males and Whites over Blacks and Hispanics) before they encountered an unexpected result (i.e, potential DIF that favors Blacks and Hispanics over Whites). They also were asked to develop explanations for the Black and Hispanic results simultaneously because educational achievement test scores of these two minority groups tend to be similar and because the DIF results in this study are identical for these two groups (see Table 3 below). Likewise, panel members were asked to develop hypothesized explanations simultaneously for both flagged explanatory steps in the White-Asian comparison.

Results

Narrative and explanatory holistic essay score means and standard deviations in the six subgroups are similar to the mean and standard deviation for all grade 9 students (narrative mean and S.D. are 3.47, 0.68; explanatory mean and S.D. are 3.26, 0.78), although they differ across subgroups. Rater agreement rates in all subgroups are similar to overall agreement rates in recent years (1989 overall exact agreement is 77 percent in narrative, 72 percent in explanatory). Finally, standardized weighted mean square fit values (see Wright & Masters, 1982, p. 101)

for both prompts in all six subgroups range from 5.7 to 7.4. These significant misfits have appeared in previous calibrations of the MWT in the PC model (e.g., Ferrara & Walker-Bartnick, 1989). This misfit is cause for some concern, although not necessarily reason to reject the PC model for these data, because large item misfit values are likely in tests with (a) small numbers of items (Wright & Masters, 1982, p. 101); (b) restricted score variability and test mistargeting (Wright & Masters, 1982, p. 101), which often evolve over time in minimum-competency tests that are closely linked to instructional programs; and because (c) calibrations of the MWT in the PC model do not unduly affect essay score equivalences and prompt difficulties (Huynh & Gleaton, 1987), as determined in comparison to results from equipercentile score equating (see Ferrara & Walker-Bartnick, 1989).

Statistical Results

Table 2 contains step difficulty values from separate calibrations of narrative and explanatory essay scores in each of the six subgroups. These step values result from calibrations on four score categories and three steps. Frequencies in score categories 1.0 through 2.5 have been collapsed to a single category because frequencies in these individual categories were low, producing extreme negative step values and high standard errors. The reader may notice that step values for each prompt within each subgroup are

monotonically increasing, as are the narrative prompt standard errors, while the explanatory prompt standard errors tend to follow a U-shape (see Table 2). Further visual inspection of the table indicates rather large step difficulty differences at some steps for Whites versus the other three race-ethnic subgroups and for males versus females.

Table 2

Step Difficulty Values for Subgroups

	Narrative			Explanatory		
Steps	1	2	3	1	2	3
White	-3.05 (.17)	.57 (.10)	1.36 (.09)	-2.04 (.13)	1.13 (.09)	2.02 (.10)
Asian	-4.40 (.29)	.46 (.12)	1.02 (.10)	-1.34 (.15)	1.34 (.10)	2.92 (.12)
Black	-3.56 (.17)	.12 (.10)	1.08 (.09)	-1.80 (.11)	1.33 (.09)	2.84 (.12)
Hispanic	-3.56 (.17)	.17 (.10)	1.08 (.09)	-1.81 (.12)	1.36 (.09)	2.76 (.11)
Male	-3.63 (.17)	.68 (.10)	1.37 (.10)	-1.85 (.12)	1.14 (.09)	2.29 (.12)
Female	-3.58 (.22)	.63 (.10)	1.00 (.09)	-2.02 (.15)	1.33 (.09)	2.64 (.11)

Note. Standard errors of calibration are in parentheses.

Figure 1 displays score Category Probability Curves (CPCs) for the explanatory prompt calibrated in the White sample. Each curve indicates the probability of an examinee at a given level of ability on the logit scale scoring in

one of the explanatory prompt score categories (i.e, 1, 2, 3, or 4). The point at which adjacent curves cross locates the step difficulty; that is, the point on the ability scale at which the probability of scoring in the higher adjacent score category becomes higher than the probability of scoring in the lower adjacent category. Difficulty values for steps 1-3 of this prompt are indicated on the figure (see Figure 1). As is typical for MWT prompts, CPCs are higher for whole number score categories (i.e., holistic scores 3.0 and 4.0, categories 2 and 4) and lower for half-point score categories (i.e., holistic score 3.5, category 3).

Figure 1 about here

Draba's statistics, calculated for each reference-focus group comparison, appear in Table 3. The White student sample was used as the reference group for comparing step difficulty values in the Black, Hispanic, and Asian focus groups; similarly, the male sample was used as the reference group for the female focus group. The null hypothesis that there is no significant difference between reference and focus group step difficulty values was rejected in eight of 24 comparisons: four times in comparisons involving narrative score steps and four times in comparisons of explanatory score steps (see Table 3).

Table 3

Draba's Statistic for Subgroup Comparisons

	Narrative			Explanatory		
Steps	1	2	3	1	2	3
Asian[a]	4.02*	0.70	2.53	-3.53*	-1.56	-5.76*
Black[a]	2.12	3.18*	2.20	-1.41	-1.57	-5.23*
Hispanic[a]	2.12	2.83*	2.20	-1.47	-1.81	-4.98*
Female[b]	-0.18	0.35	2.75*	0.88	-1.49	-2.15

Note. Negative values indicate potential bias against a focus group; positive values indicate potential bias against a reference group.

[a]Focus groups compared to Whites as reference group. [b]Focus group compared to males as reference group.

*$p<.01$, one-tailed.

DIF was detected at three steps in the White-Asian comparisons (one narrative step, two explanatory steps), the same two steps in both White-Black and White-Hispanic comparisons (one narrative step, one explanatory step), and at one step in male-female comparisons (narrative step only). In all four flagged narrative score steps, the direction of the DIF "favors" the focus group; that is, the potential "bias" is "against" Whites in the race-ethnic comparisons and "against" males in the sex comparisons. The opposite is true for the four flagged explanatory score steps: the DIF or potential "bias" is "against" the minority and female subgroups. Essay item "bias" that favors minority students over White students may be surprising;

similarly, essay item "bias" that favors males over females may be an unexpected result (see, for example, Applebee, Langer, & Mullis, 1986, p. 10).

Figure 2 displays CPCs for explanatory score categories 3 and 4 for both the White and Asian subgroups. The curves illustrate the "bias" at step 3 in the explanatory prompt (see Table 3). Figure 2 shows the between-group differences of both the (a) category 3 and category 4 curves, and (b) step 3 difficulty values (see Figure 2). These between-group differences help account for the significant DIF at step 3 in this prompt.

Figure 2 about here

Panel Review Results

Categories of hypothesized explanations for each flagged step from all reference-focus group comparisons appear in Table 4, presented in the order in which they were considered by the panel. The authors created category labels and assigned hypothesized explanations to categories; the review panel then verified and modified labels, assignments, and explanations. Comparison of the categories across subgroups indicates that explanations for the narrative step 3 "bias" that favors females over males are somewhat unique to sex group comparisons. For example,

hypothesized explanations in the "emotional difference" category re-appear only for the narrative step 2 "bias" that favors Blacks and Hispanics over Whites. Similarly, hypothesized explanations in the "assessment artificiality" explanation category re-appear only for the White-Black/Hispanic results (again, narrative step 2 and explanatory step 3). Finally, hypothesized explanations in the "language/cultural difference," "prompt-language/cultural interaction," and "rubric-rater bias" categories were used only in race-ethnic subgroup comparisons.

Table 4

Categories of Hypothesized Explanations for "Biased" Steps from all Subgroup Comparisons

Male-Female Comparisons: Narrative Step 3 "Favors" Females
. Emotional developmental differences
. Composition ability developmental differences
. Prompt topic or wording
. Artificiality of the assessment context
White-Black/Hispanic Comparisons: Explanatory Step 3 Favors Whites
. Language/cultural differences
. Interaction between prompt topic or wording and language, cultural, and SES differences
. The scoring rubric and raters' possible tendencies toward scoring bias
. Artificiality of the assessment context

(continued on next page)

Table 4 (continued)

White-Black/Hispanic Comparisons. Narrative Step 2 Favors Blacks and Hispanics
. Emotional developmental differences
. Language/cultural differences
. Interaction between prompt topic or wording and language, cultural, and SES differences
. The scoring rubric and raters' possible tendencies toward scoring bias
. Artificiality of the assessment context
White-Asian Comparisons: Explanatory Steps 1 and 3 Favor Whites
. Language/cultural differences
. Interaction between prompt topic or wording and language, cultural, and SES differences
. The scoring rubric and raters' possible tendencies toward scoring bias
White-Asian Comparisons: Narrative Step 1 Favors Asians
. Language/cultural differences
. The scoring rubric and raters' possible tendencies toward scoring bias

Note. Specific hypothesized explanations under each category above are available from the authors.

Some hypothesized explanations under these categories are simple, while others required complicated reasoning, as illustrated in Table 5 (see Table 5 below).

Table 5

Examples of Relatively Simple and Complicated Hypothesized Explanations for DIF Results

Narrative Step 3 Favors Females Over Males Explanation Category (see Table 4): Composition Ability Developmental Differences
"Girls may be better than boys at 'controlling' language."
Explanatory Step 3 Favors Whites Over Blacks/Hispanics Explanation Category (see Table 4): The Scoring Rubric and Raters' Possible Tendencies Toward Bias
"Blacks/Hispanics may be more likely to choose to write about activities that are inherently less complex and that provide less opportunity to develop detail. Raters may score complex topics (e.g., how to assemble an atom bomb) more leniently or favorably than less complex topics (e.g., how to bake cookies), even if quality of detail and control is equivalent in the two essays. In the Maryland explanatory rubric, differences in amount and quality of detail are more crucial at step 3 (i.e., the 3.5/4.0 line) than they are at lower steps (e.g., the 1.0/1.5 line)."

Discussion and Conclusions

Statistical Procedures

The statistical procedures in this study, including calibrating prompts in separate subgroups and calculating Draba's DIF statistic, are straight-forward. However, other aspects of the statistical analysis are less so. For example, the notion that DIF can occur at some score category steps but not others in the same essay prompt may seem puzzling, at least at first. In addition, the quality

of calibrations of these prompts is subject to question because of misfit of both prompts in all subgroups. Although we do not consider significant misfit statistics reason to conclude that MWT data cannot be used in the PC model, we recognize that the misfit indicates some degree of data-to-model incompatibility. Similar calibration problems are likely to arise with other polychotomously scored assessments, especially those with few exercises and in which score variability and test mistargeting evolves.

Panel Review Procedures

Like the statistical procedures, procedures for selecting panelists and conducting a review are straightforward. The four expert panelists in this study appeared to have understood enough of the presentation on the PC model, test and item bias, and the purposes and procedures for this study to be able to participate comfortably and productively in discussions of hypothesized explanations. They also reported that they enjoyed the intellectual aspects of the panel discussion. In general, the panel's hypothesized explanations are good, common-sense explanations based on members' professional experiences in teaching students the writing process and in working with students from specific race-ethnic subgroups.

The panel review process and results appear to be reasonably helpful for prompt development, but less useful

at the post-administration prompt analysis stage. This is so in this study because the panel's hypothesized explanations for DIF suggest only leads for follow-up analyses of existing or subsequently collected data. However, they do not provide justifications either for (a) accepting the indicated DIF as "true" and prohibiting a prompt from subsequent use, or (b) dismissing the indicated DIF as a false positive error or as some unexplainable and dismissible phenomenon. Other aspects of the panel's hypothesized explanations also limit their usefulness. For example, while some explanations appear to emanate from professional experience and common sense, others could be considered to be based on informal psychological theory (e.g., "Girls may be more comfortable than boys in expressing anger" from the "Emotional Developmental Differences" category in the male-female narrative result) and personally held stereotypes (e.g., "Blacks are natural story-tellers" from the "Language/Cultural Differences" category in the White-Black/Hispanic narrative result). Such explanations may not warrant follow-up analysis or action regarding a prompt. Further, some explanations could easily be applied to any score category step, not just a flagged step, which diminishes their persuasiveness as an explanation of DIF detected at a specific step.

Several other observations about the panel review procedure are related to problems typical to working with a

diverse group. For example, at times it was difficult to portray a hypothesized explanation in words because panel members were not sure or disagreed about what they meant. Also, meanings of explanation categories and assignment of explanations to categories may vary according to preferences held by the person who develops the categories and categorizes explanations. In addition, the number and usefulness of hypothesized explanations is probably a function of panel members' experience with the writing of compared subgroups, writing instruction, essay scoring procedures and criteria, previous years' essay prompts, and recent trends in writing performance. Likewise, the number of hypothesized explanations is probably a function of panel members' interest in a particular prompt or subgroup, their energy level, and whether a particular DIF result is expected or counter-intuitive. For example, in this study panel members tended to generate fewer explanations and use fewer categories on the second day and at the end of both days and for the two sets of counter-intuitive results. Finally, a panel member with an "axe to grind" could lead panel discussion in counter-productive directions. Problems like these can limit the meaningfulness and usefulness of hypothesized explanations to prompt development and revision and to reduction of DIF in essay prompts.

Recommendations for Future DIF Research and Applications

Future research on DIF in polychotomously scored tests. Future research on applying DIF detection and analysis procedures to other essay and polychotomously scored tests should address several issues. First, the incidence of DIF in other essay and polychotomously scored tests is not currently known (see Benson, 1987) and should be determined. Whether DIF would be detected in other tests, and at all or only some score steps are empirical questions, although it seems likely that results similar to these would be found in other tests. In addition, the feasibility and usefulness of the panel review procedures used for this study may or may not be feasible and worthwhile for other tests. Second, measurement researchers should continue to seek the causes of and solutions to DIF in comparisons of race-ethnic, sex, and other examinee subgroups. But they should also conduct similar analyses for other comparisons relevant to student achievement; specifically, for comparisons of schools in urban, suburban, and rural areas, of schools using different instructional approaches, and of schools with predominantly minority and majority student populations. While explanations for DIF in these comparisons might not differ from the explanations for student subgroup comparisons, actions taken in response probably would.

Third, measurement researchers should also investigate impacts on numbers of flagged steps and on the magnitude of significant DIF statistics as numbers of compared subgroups and score categories varies. It seems likely that as numbers of subgroups and score categories increase, the chance of detecting significant DIF also increases. Such a relationship may suggest that significant DIF statistics reflect measurement, calibration, and sampling error rather than a real disparity in item functioning. Huynh and Gleaton (1987) provide an adjustment to standard normal curve critical values for Draba's statistic (and Rasch model between-item fit statistics) to control for Type I error in large examinee samples. Other adjustments to statistical DIF detection procedures may also prove necessary to control for other spurious causes of statistical significance. Fourth, results from research on differences in race-ethnic and sex subgroups, in instructional approaches, and in learning and performance at different ability levels would aid understanding and explanation of DIF results, and might be useful in developing essay prompts and other open-ended exercises that are as "unbiased" as possible.

Applications in other polychotomously scored tests. Below we make several recommendations for applications of these procedures in other contexts. First, the panel review procedures we describe probably are most useful as part of the prompt development and field test process, as discussed

earlier. Of course, calibration of essay prompts and other open-ended exercises in the PC model requires larger samples (at least 400-500 examinees to achieve stable calibrations) than are often available for special field test administrations; and embedding field test prompts and exercises in operational tests is sometimes not feasible because of the time required to respond to such exercises. Second, other DIF panel reviews should include a review of student essays and responses, a step not taken in this study because of time constraints, to verify and elaborate on hypothesized explanations and guide follow-up research.

Third, test developers who detect DIF in their tests will have to decide what to do once it has been detected. Several questions must be answered in developing a policy for taking action on detected DIF. These questions include: (a) Should test developers discard, revise and re-field-test, or only monitor a flagged prompt in subsequent administrations? (b) Should test developers take action on prompts with only one flagged step or with several? How many flagged steps should require action? (c) Can test developers ignore DIF detected at some steps (e.g., steps below and above a cut score), but respond to DIF at other steps (e.g., steps adjacent to a cut score)? (d) What degree of DIF (e.g., alpha level for significance testing) can be tolerated?

Explanations of the causes and practical significance of DIF in essay prompts may aid future prompt development and identify fairness and validity issues in essay testing that may require closer scrutiny in future research. We know of no other procedures, other than judgmental reviews, proposed to detect and explain DIF in essay tests and other polychotomously scored assessments. In this paper we have described what appears to be a viable procedure for conducting such analyses. And, we hope, this study reinforces our view that concerns for fairness, reliability, and validity applied to and investigated in typical multiple-choice tests must be applied to essay tests and other polychotomously scored assessments.

References

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1986). The writing report card: Writing achievement in American schools (Report No. 15-W-02). Princeton, NJ: Educational Testing Service.

Baghi, H., & Ferrara, S. (1989). A comparison of IRT, Delta Plot, and Mantel-Haenszel techniques for detecting differential item functioning across subpopulations on the Maryland Test of Citizenship Skills. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Baghi, H., & Ferrara, S. (1990). Detecting differential item functioning using IRT and Mantel-Haenszel techniques: Implementing procedures and comparing results. Paper presented at the annual meeting of the Eastern Educational Research Association, Clearwater, Florida.

Benson, J. (1987). Detecting item bias in affective scales. Educational and Psychological Measurement, 47, 55-67.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). Assessing writing skill. New York: College Entrance Examination Board.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), Educational measurement (3rd ed.). New York: Macmillan.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rhinehart and Winston.

Ferrara, S. F. (1987). Effects of essay order on raters' score assignments. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Ferrara, S., & Walker-Bartnick, L. (1989). Constructing an essay prompt bank using the Partial Credit model. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Goldberg, G. L., & Walker-Bartnick, L. (1988). Monitoring scoring standards over a rubric transition process. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Harris, J., Laan, S., & Mossenson, L. (1988). Applying partial credit analysis to the construction of narrative writing tasks. Applied Measurement in Education, 1 (4), 335-346.

Huynh, H., & Gleaton, J. (1987). Adjustment of critical values for fit and Draba bias statistics based on the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Masters, G., & Beswick, D. (1986). The construction of tertiary entrance scores: Principles and issues. Available from the Center for the Study of Higher Education, University of Melbourne, Australia.

Phillips, G. W. (1987). Essay topic equating: Keeping the scales constant. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Phillips, G. W., Mead, R. J., & Ryan, J.P. (1983). Use of the general polychotomous form of the Rasch model as a means of calibrating and equating indirect and direct assessments of writing achievement. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. Language Testing, 4 (1), 72-92.

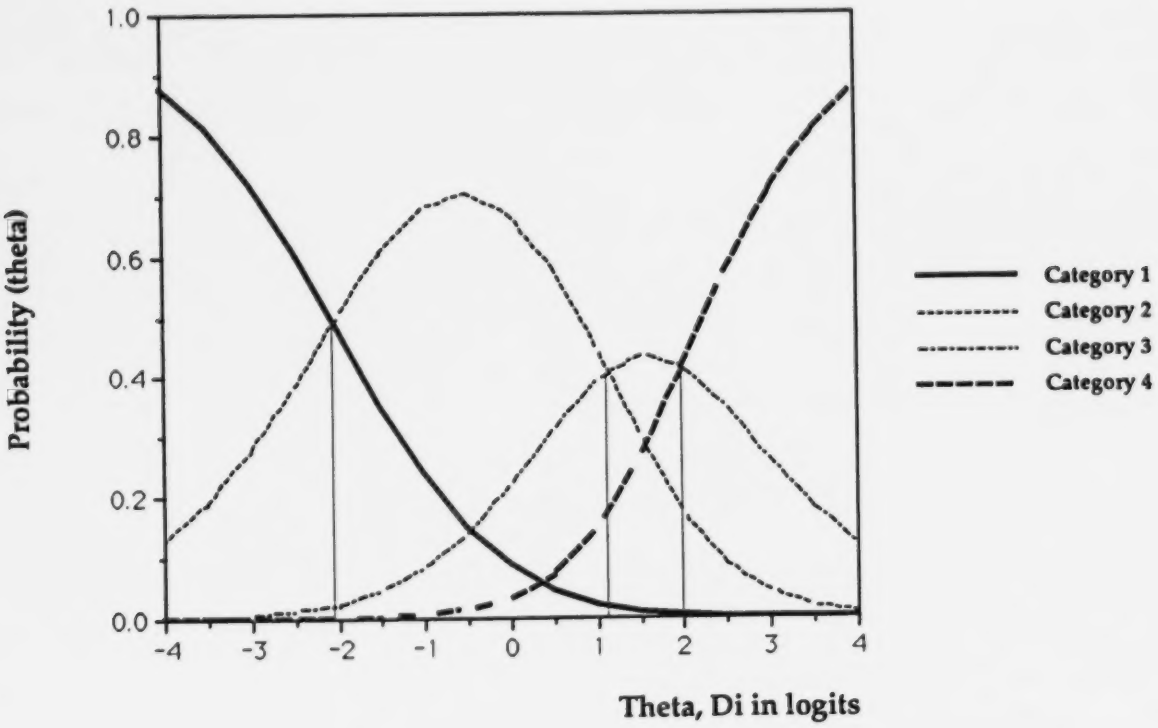
Wright, B. D., Congdon, R., & Schultz, M. (1988). MSTEPS [Computer program]. Chicago: University of Chicago.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago: MESA Press.

sf1608

Figure 1

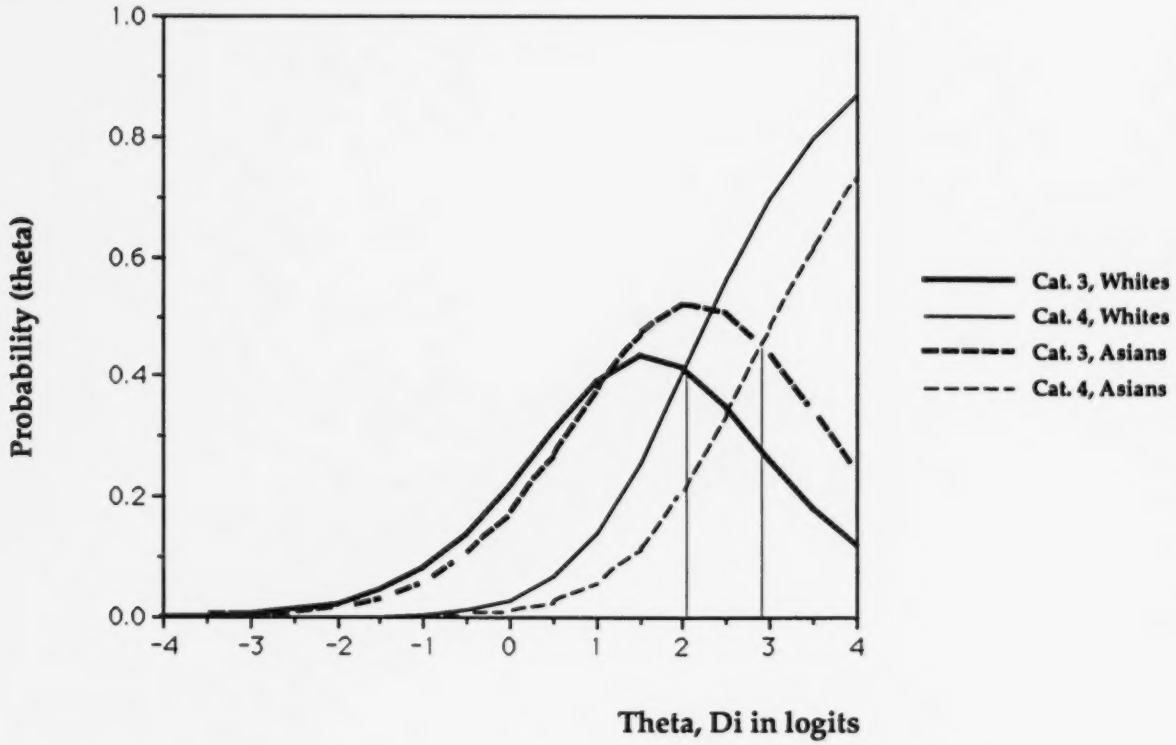
Category Probability Curves for Whites: Explanatory Prompt



Note: Di=step difficulty. Vertical lines indicate step 1, 2, and 3 difficulty values (see Table 2).

Figure 2

Category Probability Curves for Explanatory Score Categories 3 and 4, White and Asian Subsamples



Note: **Di=step difficulty. Vertical lines indicate White step 3 difficulty (2.02) and Asian step 3 difficulty (2.92).**